Help International - Clustering

Rahul Raj - 5rd, Sept 2022

Problem Statement

HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.

After the recent funding programmes, they have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.

Business Objective

Our job is to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then you need to suggest the countries which the CEO needs to focus on the most. The datasets containing those socio-economic factors and the corresponding data dictionary are provided below.

DataSet

Country-data.csv

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553
1	Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090
2	Algeria	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460
3	Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530
4	Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200

** Shape of the dataset : 167 rows , 10 columns

Correlation between the Original variables

Top 5 Correlations

	var_1	var_2	abs_corr_value	corr_dir
1	income	gdpp	0.895571	Positive
2	child_mort	life_expec	0.886676	Negative
3	child_mort	total_fer	0.848478	Positive
4	life_expec	total_fer	0.760875	Negative
5	exports	imports	0.737381	Positive

`income` & `gdpp` are highly correlated by `0.89` `Positively`

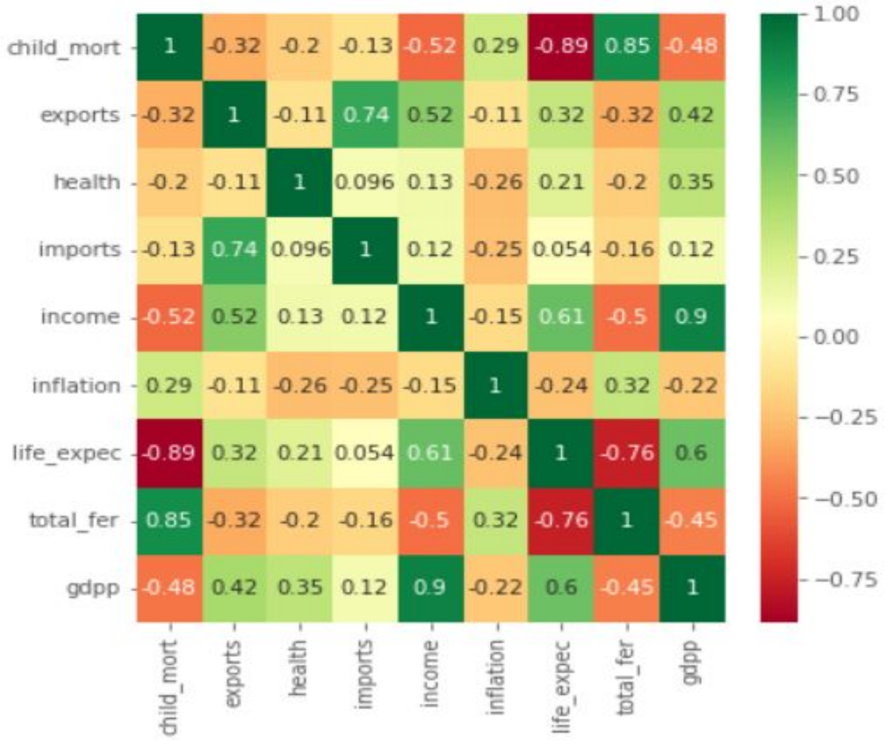
`child_mort` & `life_expec` are `Negatively` correlated by `0.88`

`child_mort` & `total_fer` are `Positively` correlated `0.85`

`life_expec` & `total_fer` are `Negatively` correlated by `0.76`

`exports` & `imports` are `Positively` correlated by `0.73`

Correlations Matrix



Principal Component Analysis (PCA)

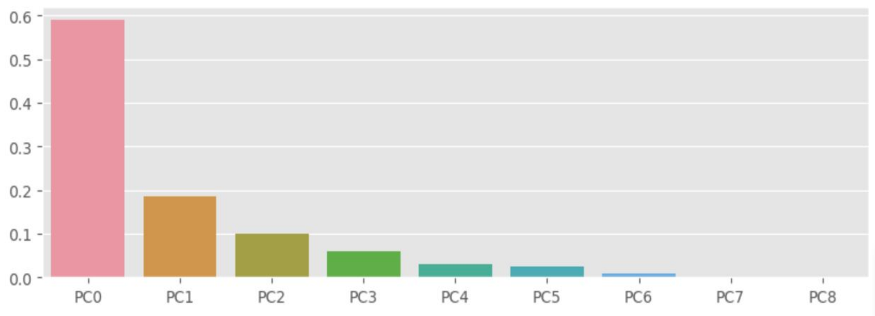
`PCA` helps remove the redundancies in the data and find the most important directions where the data was aligned.

Principal component analysis (PCA) is one of the most commonly used `dimensionality reduction` techniques in the industry.

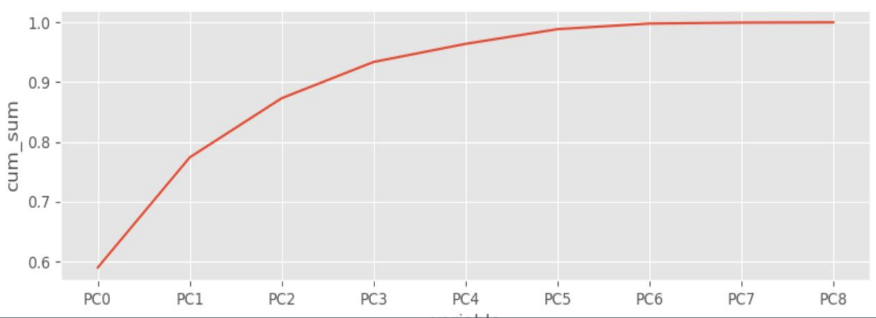
By converting large data sets into smaller ones containing fewer vaPCA Component Cumulative Varianceriables,

it helps in `improving model performance`, `visualising complex data sets`, and inPCA Component Cumulative Variance many more areas.

PCA Component Variance



PCA Component Cumulative Variance

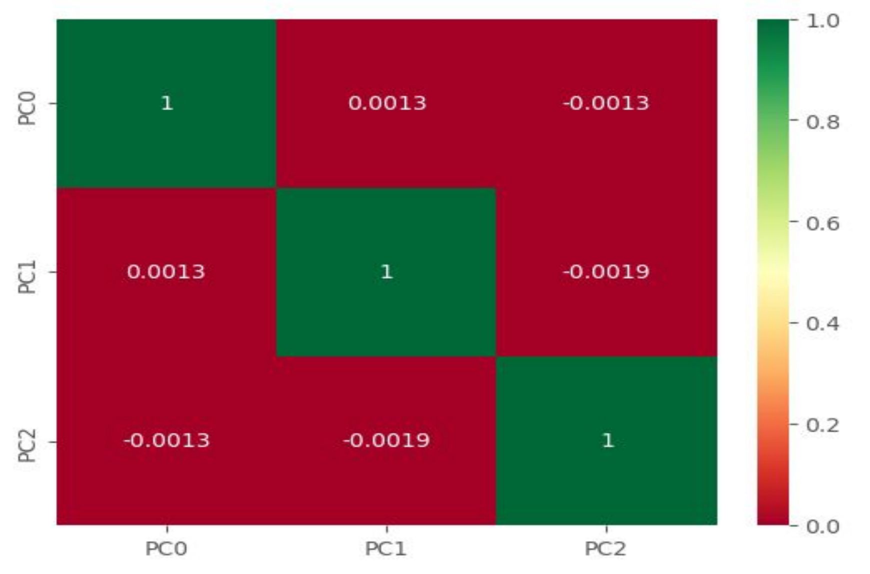


The combination of first `3 components` explains about `87%` of the variance in the data

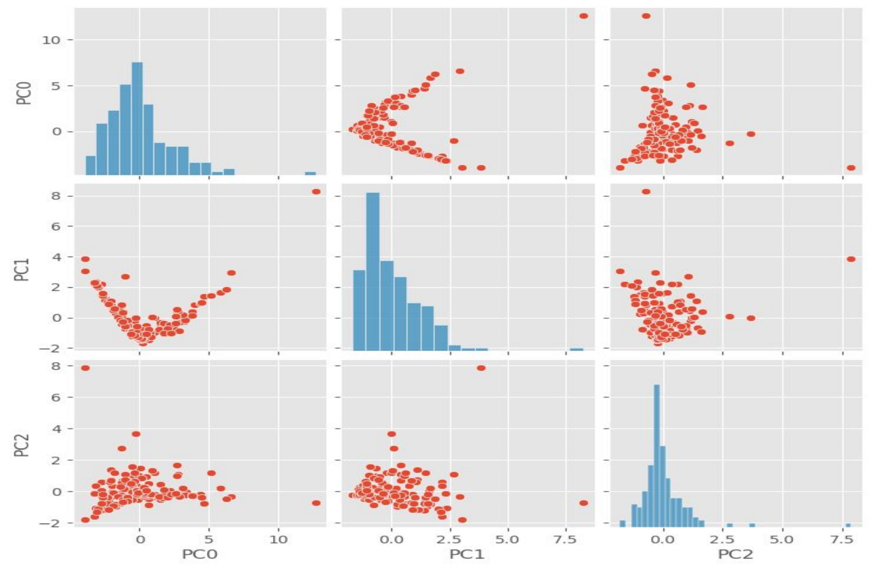
Hence, we will use these components only, for further processing

Correlation of Principal Components

Principal Components Correlation Heat Map



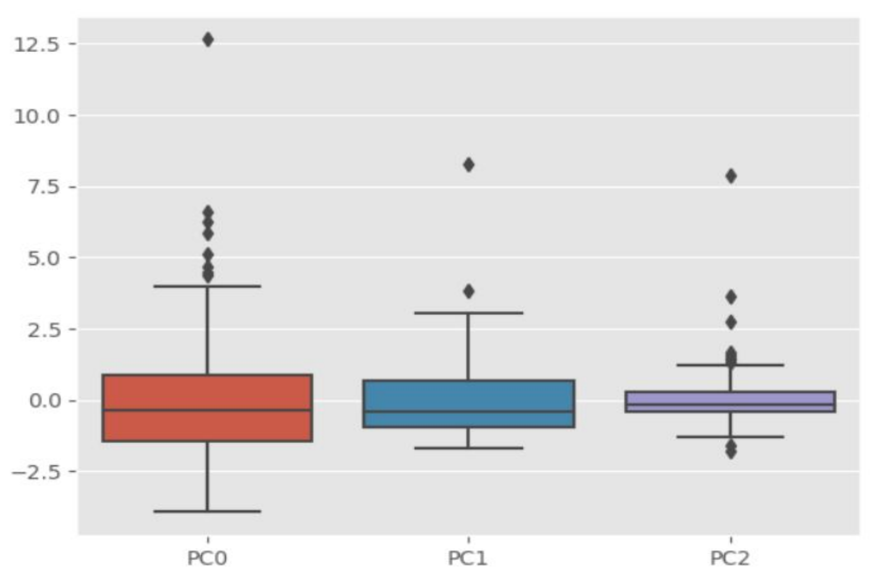
Principal Components Relations on Scatter Plot



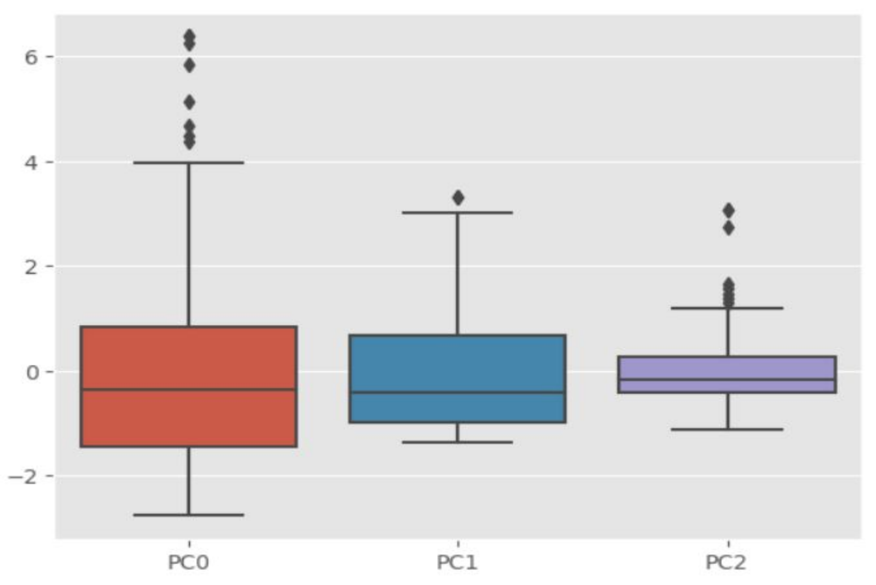
There is no correlation between any Principal components, which is very good

Treatment of Statistical Outliers

Before Treatment



After Treatment



We choose `Capping` the outliers with Q1 & Q3 values,

Because we don't want any countries be removed from the dataframe and we need all the countries needs to be clustered as well

We took Q1 as 5th Percentile and Q3 as 99th Percentile

Model Building

Clustering Model

For Model Building

KMeans Clustering

Hierarchical Clustering

To Find the Optimum K

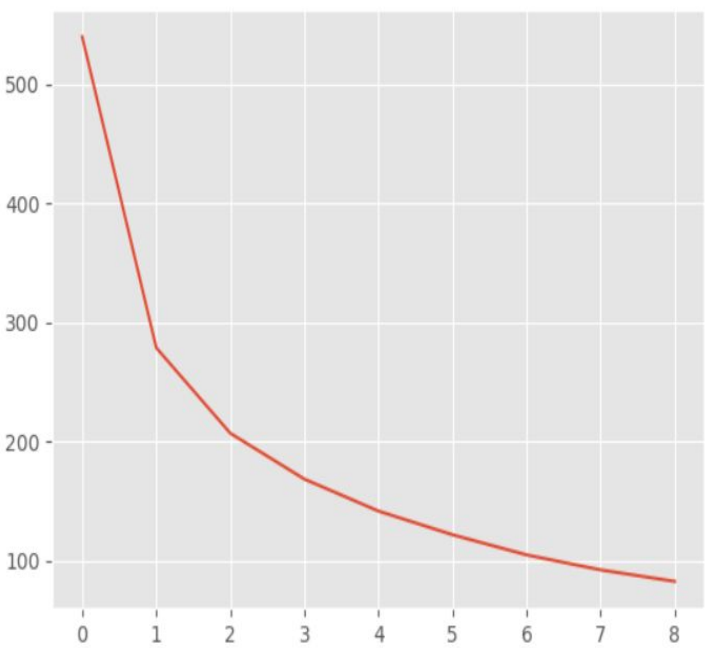
Elbow Curve

Silhoutte Analysis

K's with ssd & silhouette score

	k	ssd	silh
0	2	539.55	0.50
1	3	278.58	0.49
2	4	206.85	0.48
3	5	168.37	0.40
4	6	141.54	0.42
5	7	121.68	0.38
6	8	104.77	0.39
7	9	92.18	0.37
8	10	82.54	0.37

Elbow Curve

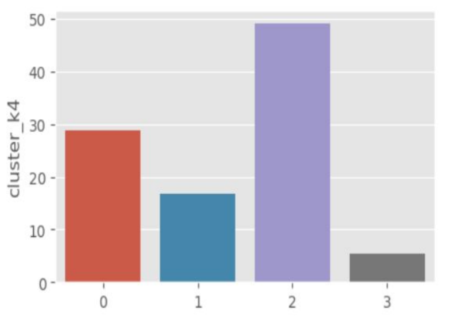


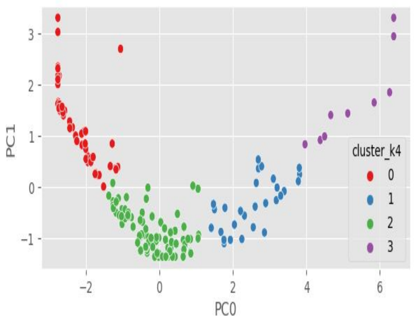
- Looking at the `Elbow curve` (ssd) and `silhouette score`, `4` or `5` clusters could be ideal in this problem

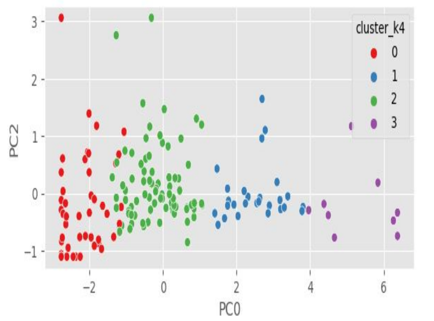
KMeans Clustering

K = 4

	cluster_k4	count	%distri.
0	2	82	49.0
1	0	48	29.0
2	1	28	17.0
3	3	9	5.0

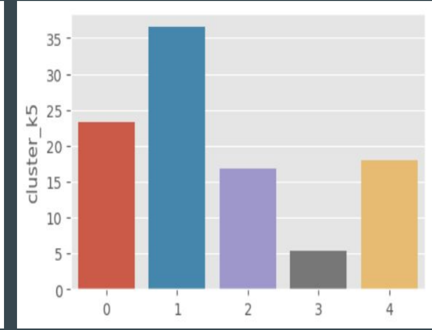


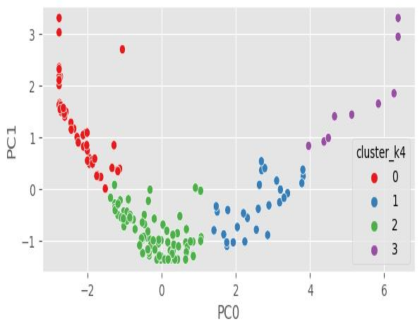


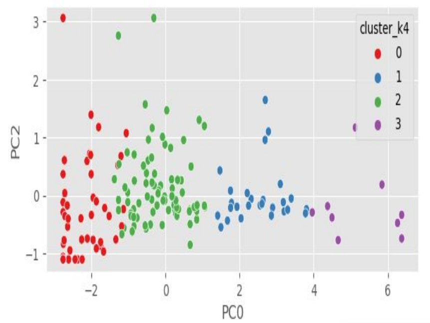


K = 5

	cluster_k5	count	%distri.
0	1	61	37.0
1	0	39	23.0
2	4	30	18.0
3	2	28	17.0
4	3	9	5.0



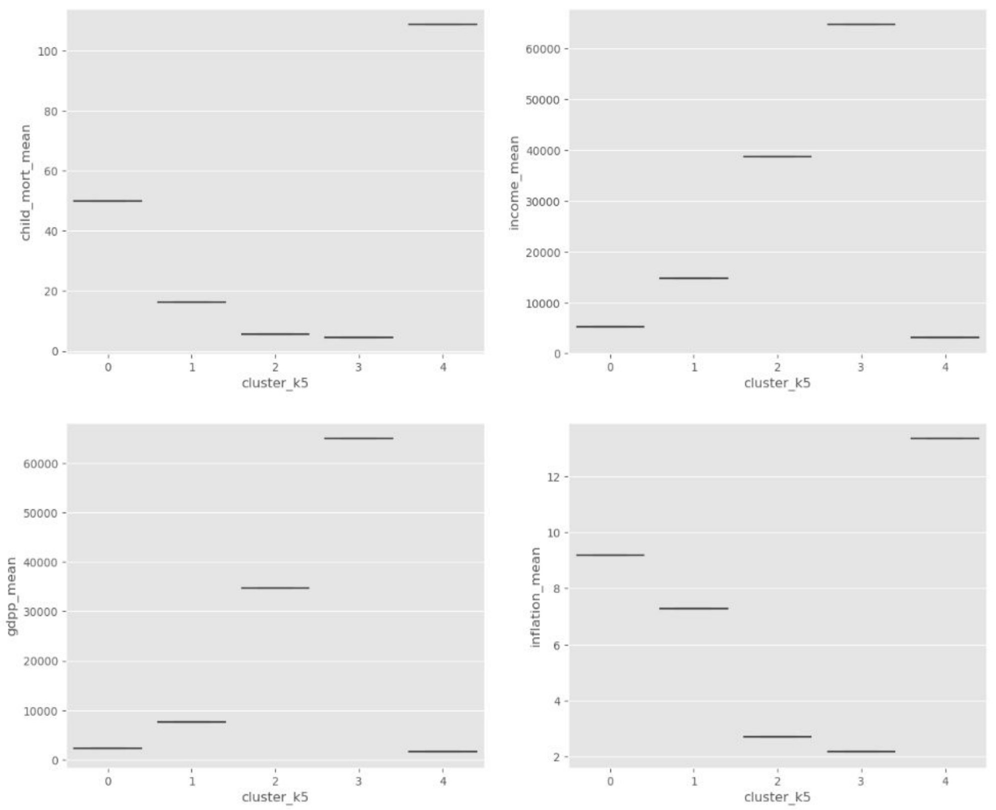




K = 5 , is the better model

Mean Value of Clusters, K = 5

Mean Value on Plot



Mean Value Table

	cluster_k5	child_mort_mean	income_mean	gdpp_mean	inflation_mean
0	0	50.0	5193.0	2360.0	9.0
1	1	16.0	14775.0	7686.0	7.0
2	2	6.0	38721.0	34718.0	3.0
3	3	4.0	64767.0	65078.0	2.0
4	4	109.0	3077.0	1544.0	13.0

Cluster 4 & Cluster 0 can be considered for our further Analysis

Cluster 4

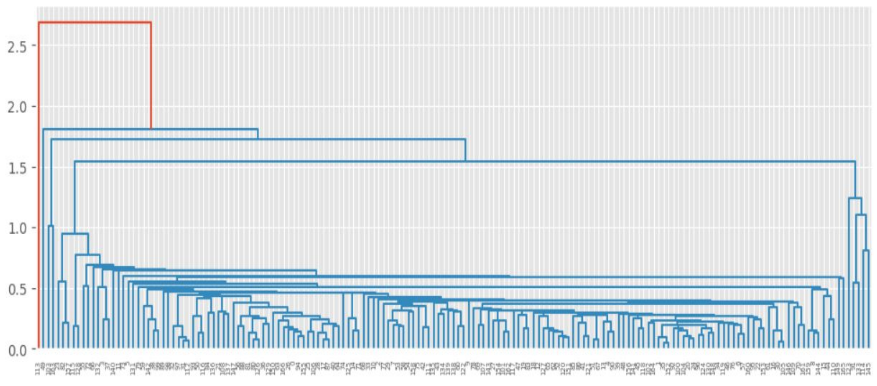
- higher child mortality
- low income
- low gdpp
- high inflation

Cluster 0

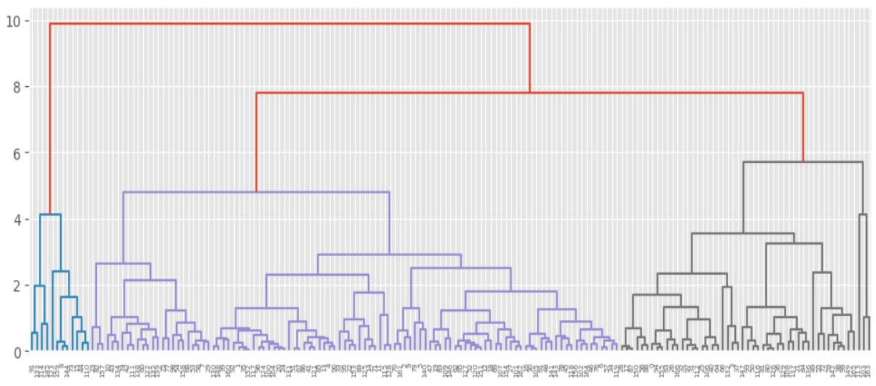
- moderate child mortality
- low income
- low gdpp
- moderate inflation level

Hierarchical Clustering

Single Linkage

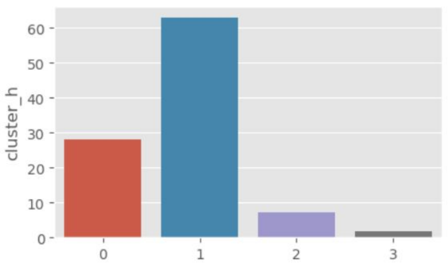


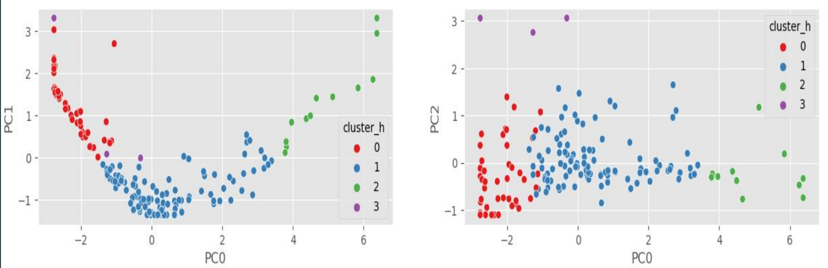
Complete Linkage



We cut tree @ n_clusters = 4

	cluster_h	count	%distri.
0	1	105	63.0
1	0	47	28.0
2	2	12	7.0
3	3	3	2.0





Choosing the Best Model & Deducing the List

- We have analyzed both `K-means` and `Hierarchical clustering` and found clusters formed are not identical.

- The clusters formed in both the cases are not that great but `it's better in K-means` as compared to Hierarchical.

- Hence, we will proceed with the clusters formed by `K-means with k = 5` and based on the information provided by the final clusters**

- We will `deduce the final list of countries` which are in need of aid.

<u>- It is obvious that we need to choose `cluster 4` & `cluster 0` are the be considered for further proceedings</u>

	count	mean	std	min	25%	50%	75%	max
child_mort	69.0	76.0	38.0	17.0	47.0	64.0	100.0	208.0
exports	69.0	862.0	1921.0	1.0	110.0	305.0	730.0	14672.0
health	69.0	116.0	146.0	13.0	37.0	57.0	121.0	766.0
imports	69.0	904.0	1359.0	1.0	215.0	428.0	1182.0	10072.0
income	69.0	4273.0	4869.0	609.0	1540.0	2660.0	5190.0	33700.0
inflation	69.0	11.0	14.0	1.0	4.0	8.0	15.0	104.0
life_expec	69.0	62.0	7.0	32.0	58.0	62.0	67.0	72.0
total_fer	69.0	4.0	1.0	2.0	3.0	5.0	5.0	7.0
gdpp	69.0	2005.0	2518.0	231.0	595.0	1170.0	2740.0	17100.0
PC0	69.0	-2.0	1.0	-3.0	-3.0	-2.0	-1.0	-1.0
PC1	69.0	1.0	1.0	-1.0	-0.0	1.0	1.0	3.0
PC2	69.0	-0.0	1.0	-1.0	-1.0	-0.0	0.0	3.0
cluster_k4	69.0	1.0	1.0	0.0	0.0	0.0	2.0	2.0
cluster_k5	69.0	2.0	2.0	0.0	0.0	0.0	4.0	4.0

Deduced List of countries

Since the list is a bit longer,
We choose the following variables

- `child_mort`
- `income`

for deducing with list based on the mean values of listed countries

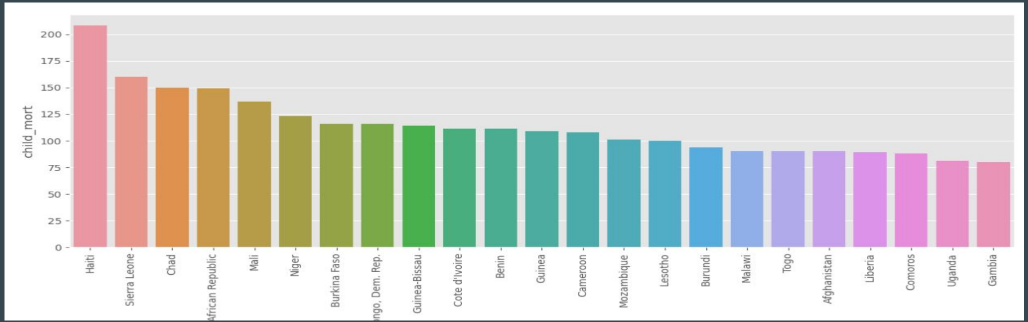
	count	mean	std	min	25%	50%	75%	max
child_mort	23.0	114.0	30.0	80.0	90.0	109.0	120.0	208.0
exports	23.0	165.0	141.0	21.0	79.0	127.0	188.0	617.0
health	23.0	42.0	23.0	18.0	31.0	37.0	46.0	130.0
imports	23.0	293.0	223.0	91.0	170.0	248.0	328.0	1182.0
income	23.0	1445.0	588.0	609.0	974.0	1410.0	1740.0	2690.0
inflation	23.0	7.0	5.0	1.0	3.0	5.0	10.0	21.0
life_expec	23.0	56.0	7.0	32.0	55.0	57.0	59.0	66.0
total_fer	23.0	5.0	1.0	3.0	5.0	5.0	6.0	7.0
gdpp	23.0	627.0	289.0	231.0	432.0	562.0	733.0	1310.0
PC0	23.0	-3.0	0.0	-3.0	-3.0	-3.0	-2.0	-2.0
PC1	23.0	2.0	1.0	1.0	1.0	1.0	2.0	3.0
PC2	23.0	-1.0	0.0	-1.0	-1.0	-1.0	-0.0	0.0
cluster_k4	23.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
cluster_k5	23.0	4.0	1.0	0.0	4.0	4.0	4.0	4.0

FINAL LIST OF COUNTRIES

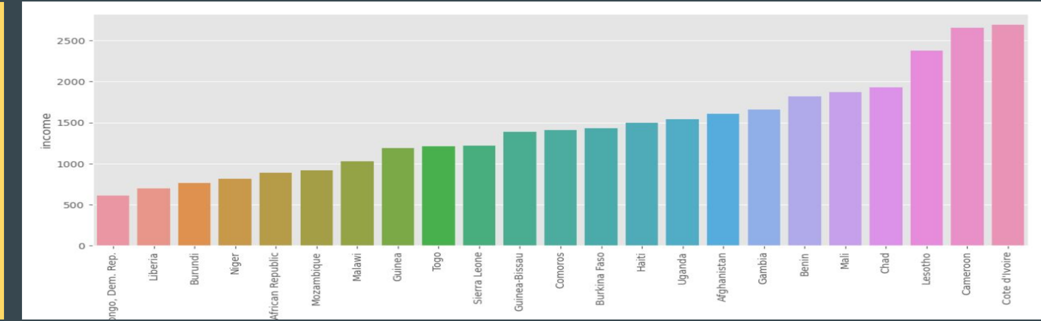
Total 23 Countries

0	Afghanistan
1	Benin
2	Burkina Faso
3	Burundi
4	Cameroon
5	Central African Republic
6	Chad
7	Congo, Dem. Rep.
8	Cote d'Ivoire
9	Gambia
10	Guinea
11	Guinea-Bissau
12	Haiti
13	Lesotho
14	Liberia
15	Malawi
16	Mali
17	Mozambique
18	Niger
19	Sierra Leone
20	Togo
21	Uganda
22	Comoros

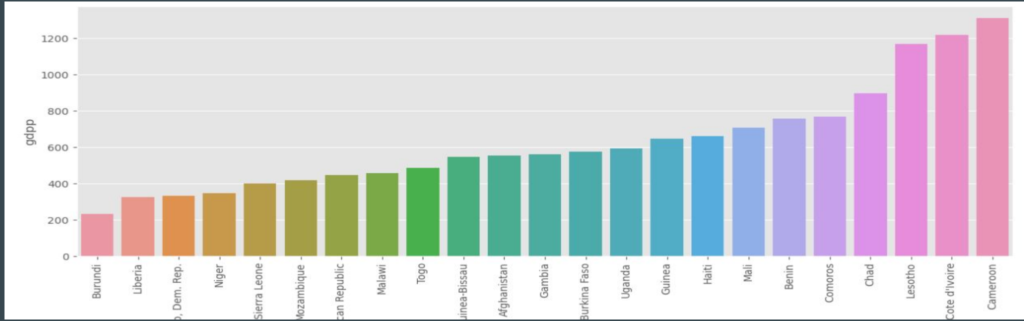
CHLD MORTALITY BY COUNTRY



INCOME BY COUNTRY



INCOME BY COUNTRY



End of the Report